DOCUMENT RESUME

ED 279 658 SP 028 683

AUTHOR Konrad, Abram G.
TITLE Student Evaluation of Teaching: Research and Recommendations.
PUB DATE Jun 86
NOTE 45p.; Paper prepared for the Committee for the Improvement of Teaching and Learning, the University of Alberta.
PUB TYPE Information Analyses (070)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *College Faculty; *Educational Assessment; Evaluation Criteria; Higher Education; Instructional Improvement; Program Effectiveness; *Student Evaluation of Teacher Performance; *Teacher Effectiveness; Teacher Improvement

ABSTRACT

This paper provides an overview of instructional evaluation in higher education with a special emphasis upon student evaluation of teaching. The paper includes an examination of the context for evaluation, components of evaluation, and research on student evaluation of instruction, and it proposes a series of recommendations for the implementation of student evaluations. A major portion of the paper reviews selected research literature regarding such matters as the validity, reliability, influential factors, instrumentation, administration, and impact of student evaluation of teaching. Research evidence on the impact of student ratings on the improvement of teaching is also examined. The kind of rating scales that should be used and how rating forms should be administered are considered. An eight-page reference list concludes the document. (JD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED279658

STUDENT EVALUATION OF TEACHING : RESEARCH AND RECOMMENDATIONS

by
Abram G. Konrad

Prepared for
The Committee for the Improvement of Teaching and Learning
The University of Alberta
June, 1986

SP028683

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

A. G. Konrad

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Executive Summary

The evaluation of university education continues to attract considerable attention within higher education. At the University of Alberta, the Committee for the Improvement of Teaching and Learning has undertaken the development of a computerized system of student evaluation of instruction. As part of this development, the Committee commissioned a review of relevant literature in instructional evaluation.

This paper provides an overview of instructional evaluation in higher education, with a special emphasis upon student evaluation of teaching. The paper includes an examination of the context for evaluation, components of evaluation, and research on student evaluation of instruction, and it proposes a series of recommendations for the implementation of student evaluations. A major portion of the paper reviews selected research literature regarding such matters as validity, reliability, influential factors, instrumentation, administration, and impact of student evaluation of teaching.

Although the literature clearly indicates that there is no one way to evaluate instruction, institutions would benefit from adopting a systematic approach to instructional evaluation. The following recommendations were derived from the literature on the evaluation of university teaching:

1. A multifaceted system of instructional evaluation should be adopted, including student, peer, administrator and self-evaluation activities.

2. The evaluation system must be both credible and visible.

3. Student rating forms should be adopted or adapted to provide information for improved decision making in personnel matters as well as for instructional improvement.

4. Standard procedures for conducting student ratings should be adopted.

5. Avoid rating fatigue; not all courses should be evaluated each term.

6. The student rating form should provide for flexibility by allowing instructors to develop optional items and inviting students to write comments in response to open-ended questions.

7. A small number of three to five global items could be used to measure the overall effectiveness of teaching for personnel decisions.

8. Diagnostic items should be developed for each of the major dimensions of teaching effectiveness and used to provide information to the instructor for teaching improvement.

9. The confidentiality of data must at all times remain inviolate

10.Evaluation results must be carefully interpreted, particularly if they are used in personnel decisions. Whenever norms are compiled on items that involve sufficient responses, they should reflect the influence of extraneous factors on the evaluation results -- department, class size, level and rank.

11. The administration must actively support the policies, procedures and cost of the evaluation system.

12. There should be a clear linkage between evaluation and the reward system of the university.

13. Data storage security must be guaranteed.

14. Adequate time should be provided for the development, acceptance and implementation of student evaluation policies and procedures.

15. The student evaluation system should be staffed by an expert in instructional evaluation directly accountable to the academic vice-president, and the system should be monitored by a faculty-student advisory committee.

Table of Contents

Section	Page
Context of Evaluation	1
Components of Evaluation	2
Purposes	2
Areas	4
Elements	5
Procedures	7
Research on Student Evaluation	11
1. How valid are student evaluations?	14
2. How reliable are student ratings?	16
3. What factors influence student ratings?	17
4. What is the impact of student ratings?	21
5. What kind of rating scales should be used?	23
6. How should rating forms be administered?	27
Recommendations	29
References	33

STUDENT EVALUATION OF TEACHING: RESEARCH AND RECOMMENDATION

Abram G. Konrad

The concern for educational excellence has recently focused upon instructional evaluation. The bastion of academic freedom that established the classroom as a professor's preserve has crumbled to the outcry of the public for accountability and efficiency, and to the students' expectations for a quality education that guarantees employability. University teaching can no longer enjoy immunity from careful scrutiny. Those committed to standards of excellence and who seek instructional improvement herald this development with enthusiasm.

Evaluation means different things to different people. For instance, it is describing a program or person as fully as possible (Stake, 1967); it is providing useful information for decision makers (Stufflebeam, 1969); it is both the judgment on the worth or impact of a program, procedure, or individual, and the process by which that judgment is made for decision making (Dressel,1976). To evaluate,accordingly, is to place a value of worth on something or someone for the purpose of decision making or change.

Not all professors respond favourably to the notion of instructional evaluation. Often, professors resist evaluation because of a fear that the results will be misused. Administrators are viewed with suspicion when they encourage the adoption of evaluation procedures that serve primarily the interests of the institution. Students are criticized for publishing "course guides" based upon instructional evaluation that appears to be little more than a popularity poll. Marsh (1984) contends that these ill-founded fears persist, in part, because faculty members frequently have no teacher training, and yet their salaries and even their

jobs depend upon their instructional abilities. And yet, any professor, or instructional unit, for that matter, that wishes to achieve greater excellence in teaching must engage in some form of instructional assessment.

Evaluation of the quality of postsecondary teachers and teaching has been debated extensively over the past 20 years. In a comprehensive bibliography on instructional evaluation five years ago, Murray (1980) claimed the number of publications on teaching evaluation exceeded 2000. A report, "Procedures for Evaluating Instruction in a University Setting," prepared by the Committee for the Improvement of Teaching and Learning at the University of Alberta (Monod et al., 1979), provided a broad overview of major evaluation processes and presented a series of recommendations to the General Faculties Council of the University. This paper draws upon selected research literature on instructional evaluation and presents a series of recommendations or guidelines for the development of strategies for student evaluation of teaching at the University of Alberta.

Components. As an outcome of its study of evaluation practices among colleges and universities, the Southern Regional Education Board (1977) developed a framework for a systematic faculty evaluation program. The framework identifies four components common to all evaluation programs: purposes, areas, elements and procedures.

Purposes. Why Evaluate? Not all professors respond favorably to the notion of instructional evaluation; but how can it be avoided? Faculty members can no longer remain aloof to quality assessment. Among the purposes served by instructional evaluation, the following seem to be the most common.

1. To provide data for personnel decisions relating to faculty tenure, promotion and salary. Evaluators refer to such evaluation as summative evaluation because, as Centra (1979:16) points out, "it 'sums up' performance at the end of a time period and results in some kind of overall judgment." Data gathered for summative evaluation lead to a decision about a person or program.

Faculty members readily understand the nature of summative evaluation. They evaluate students and assign final grades as a regular part of their teaching. Faculty salary, selection, promotion and tenure decisions all involve summative evaluation.Unfortunately, experiences in personnel matters sometimes make faculty members skeptical of evaluation in higher education. Some fear that the measures used in evaluation will not treat them fairly. To be fair, personnel decisions must rely upon measures by which faculty members can be compared across a variety of instructional settings and disciplines. At the same time, data should also be collected under similar conditions and with comparable procedures.

2. To provide diagnostic feedback to help faculty members improve their performance. Evaluation designed to improve performance is called formative evaluation because "it is meant to help 'form' performance while it is in progress" (Centra, 1979:16). Formative evaluation concerns itself with improvement, development and growth.

In formative evaluation, data are gathered to identify instructional strengths and weaknesses. Such evaluation data appropriately belong to the faculty member being evaluated, and these data should be helpful in identifying areas in need of improvement. Generally, instructional improvement will result most readily when data are gathered in a non-threatening manner. To achieve greater effectiveness, a faculty member may wish to share the evaluation results with persons equipped to

provide professional assistance in specific skill areas.

Although instructional evaluation serves primarily summative and formative purposes Genova et al. (1976:8-10) identify several additional purposes. These purposes may provide further objectives in developing evaluation programs.

3. *Guiding students in their selection of courses and instructors.*

4. *Evaluating courses, curricular sequences, programs, departments and other units.*

5. *Keeping an inventory of personnel resources for reassignment and retraining within the institution.*

6. *Conducting research on factors related to teaching effectiveness.*

7. *Informing external audiences on faculty performances.*

Areas. Evaluation areas identify those aspects of faculty activities that are to be examined in determining faculty effectiveness. Although this paper focuses specifically upon instructional evaluation, other evaluation areas are also important in assessing faculty effectiveness. In its survey of evaluation practices, the Southern Regional Education Board (1977) identified nine evaluation areas in making personnel decisions in postsecondary institutions. Instructional activities were ranked as the most important area for faculty evaluation in all types of institutions. Other areas were ranked differently by different types of institutions. For example, doctoral institutions ranked research and publications second and third, and student advising fourth. It is important to specify areas of performance for faculty evaluation. When multiple areas are specified in a summative evaluation, the weight assigned to each area should be clearly established for the development of a composite score. The performance areas and their weights

should be used to reflect accurately the specific responsibilities of each faculty member.

Elements. In any evaluation, there must be criteria, standards and evidence. Criteria identify the specific behaviors or role expectations in each evaluation area. What behaviors, characteristics, or instructional aspects that need to be examined in an evaluation constitute the criteria for evaluation? These criteria may differ from one institution to another, but they should be clearly specified in the development of an evaluation program.

Meeth (1976) proposed three categories of criteria for evaluating the area of instruction -- ultimate, intermediate and immediate. The ultimate criteria focus upon instructional outcomes, including such matters as attainment of instructional goals, attainment of student goals, student learning gains, etc. Intermediate criteria are based on the process or methodology of teaching -- course organization, presentation, motivation, evaluation, etc. The immediate criteria focus on the sense of the learning experience, particularly the level of student satisfaction. For each criterion in an area of evaluation, standards of performance must be established. Standards specify levels of expectation in behaviors, characteristics or instructional aspects. Thus, a standard allows one to distinguish between an excellent and a poor teacher.

Evaluators specify two kinds of standards: norm-referenced and criterion-referenced. Norm-referenced standards compare the performance of one faculty member against the performance of colleagues. In an institution where a reward for teaching excellence is given to the best instructor, for example, evaluations are norm-referenced. Criterion-referenced standards are levels of performance established independently of the group being evaluated. Thus, an

institution could give teaching excellence awards to as many instructors as meet a predetermined level of performance.

Setting standards is a very difficult task, but it is essential in any systematic evaluation program. In particular, evaluation that leads to decision making relies upon the explication of measurement standards for each criterion.

What kind of data or information can be gathered as evidence to measure teaching effectiveness? In many institutions, systematic data collection procedures have been established; in others, evidence of teaching effectiveness is largely hearsay and impressionistic.

Evaluation data may be quantitative (descriptive) or qualitative (judgmental). A quantitative statement presents descriptive data without attaching any judgment to it. For example, "My instructor missed three class sessions." Qualitative data imply a value judgment, as in "My instructor missed too many class sessions." Many of the procedures used in instructional evaluation depend heavily upon perceptual judgments. It should be acknowledged that perceptions in themselves are neither right nor wrong; they are the views of reality held by observers or participants in a situation.

Berk (1979) synthesized research on the construction of rating scales for faculty evaluations. He found that the summated ratings technique or Likert-type scale was the most widely used approach in evaluation. A Likert-type scale can be used to rate instructional characteristics on a quality scale from poor to excellent, on a frequency scale from never to always, or on an intensity scale from little to much.

One of the best ways to ensure the validity of instructional evaluation is through gathering evidence on any given criterion from more than one source. A

good evaluation scheme is multifaceted. In the following section, the major data sources for instructional evaluation will be identified.

Procedures. The final component in a systematic evaluation program focuses upon evaluation procedures. Procedures must be developed for gathering evidence in applying the standards to the appropriate criteria for each of the evaluative purposes. In a systematic approach to instructional evaluation, a flow chart should be developed to indicate who does what, when and how. Some instructional activities should be monitored continuously, while others should only be reviewed periodically. Similarly, all classes of some faculty members (e.g.,non-tenured) should be evaluated annually, while only selected classes, including "new" courses, of other faculty members should be included in annual evaluations.

Evaluation procedures are designed to establish what is good teaching. Although the complexity of teaching is readily acknowledged, the development of evaluation procedures rests upon the assumption that teaching effectiveness can be measured. In examining a variety of measurement approaches, it should be recognized that instructional methods also vary considerably. An attempt to match certain teaching methods to academic discipline reported by Donald and Shore (1977), resulted in a taxonomy of eight teaching methods that can be placed on a continuum from teacher-centered to student-centered activities. Evaluation procedures could be developed for a particular method of instruction, but this paper focuses upon instructional evaluation in general terms.

The data sources for instructional evaluation examined here include student learning, student ratings, colleague ratings, administrator evaluations, and self evaluation.

1. Using student learning as a measure of teaching effectiveness seems to be most appropriate. What students have learned and how well they have learned it constitute two major concerns in instructional evaluation. Unfortunately, as Centra (1979) pointed out, the practical and psychometric problems make interpretation of student learning difficult. Many factors other than teacher competence can affect student "gain" scores in course examinations. Careful measurement controls would have to be devised to determine the direct impact of teaching behavior on student learning, and most faculty members (and students) would resist such constraints.

Some institutions use criterion-referenced measurement of learning outcomes,particularly in competency-based or mastery learning situations. Moomaw (1977) found this approach used more often in the performing arts and in vocational/technical programs than in the more traditional academic programs. In some institutions faculty members are rewarded according to the success rate of their students, but such criteria are subject to severe misuse. Many uncontrolled factors influence student learning.

Not only are learning outcomes difficult to relate to teaching behavior they rarely are accepted as evidence in support of teaching effectiveness. Donald and Shore (1977:69) advanced three avenues open to instructors wishing to provide such evidence of good teaching:

> . . . the description of changes made to respond to learning difficulties encountered by students; . . . the eventual success of such students; and . . . selection and design of the curriculum and reward system which most appropriately suits the students.

Evidence with regard to the learning success of students, according to Donald and Shore, should not be used in a normative process of evaluation. To measure student learning is probably the ideal method of assessing instruction, but it is also the most difficult. Its primary value as a source of information seems to lie in modifying instructional procedures rather than in making summative decisions.

2. Among the most widely used evaluation procedures are student ratings. In Canada, Knapper (1972) reported that 31 out of 36 institutions used some form of student evaluations. In the southern states, 88 percent of the institutions surveyed reported using information from students (Southern Regional Education Board, 1977). In spite of widespread use of student ratings in instructional evaluation, considerable variations exist in the content and administration of such forms. Some institutions have systematically developed forms and established procedures for common usage; at other institutions, however, faculty members design their own forms and administer them quite independently from any overall concerns for institutional evaluation.

Nadeau (1975:75) asserts that because "students are the primary agents in the learning process, "they have a "right" to participate in instructional evaluation. From student feedback, an instructor can identify relative strengths and weaknesses in his or her teaching. From a comprehensive review of research on evaluation of teaching, Marsh (1984) concludes that student evaluations are the most thoroughly studied of all forms of personnel evaluation, and one of the best in terms of being supported by empirical research. Some of this research will be examined in the next major section of this paper.

3. What is the place of colleague ratings in instructional evaluation?

Peer evaluation has a long standing tradition in colleges and universities. Usually colleagues express judgments on their peers in tenure, promotion and salary decisions. Although structures vary considerably across postsecondary institutions, it seems customary to have colleagues participate in personnel decisions through formal committee hearings.

The part that colleagues play in instructional evaluation *per se,* however, is much less clear. In the tradition of academic freedom and collegiality, the university classroom is the professor's preserve -- it is a sanctuary within which one may freely profess one's thoughts and theories without fear of sanctions. The survey of the Southern Regional Education Board (1977) found strong resistance to classroom visitation among faculty members. Visitation was perceived to threaten faculty rights. That colleagues participate in personnel decisions through committee activities is one thing, but to have them engage in a systematic approach to data collection for peer evaluation is quite another.

Batista (1976) identified a list of ten faculty behaviors most suited to peer evaluation. Among these were the following behaviors related to instructional evaluation: up-to-date knowledge of subject matter; knowledge of what must be taught; knowledge and application of the most appropriate or most adequate methodology for teaching specific content area; knowledge and application of adequate evaluation techniques for the objectives of his/her course(s); professional behavior according to current ethical standards; personal and professional attributes; and attitudes toward and commitment to colleagues, students, and the institution. Centra (1979) suggested that colleagues might better judge the substance of teaching than classroom activities by examining choice of text, course syllabus and objectives, reading lists and materials

employed in instruction, assignments and examinations, and similar indicators of substance. Such judgments could rely on data gathered through collegial interaction and participation in projects and activities.

The Southern Regional Education Board (1977) survey found that institutions which assign a greater role to faculty in the evaluation process tend to have higher faculty morale and greater acceptance of evaluation and its results than institutions which do not include as much faculty participation. The issue, then, is not whether to include peer evaluation, but how and for what purpose. Systematic procedures can improve the nature as well as the quality of colleague ratings. Classroom visits tend to be decisive and threatening, and should be used for formative rather than summative purposes. In a supportive collegial climate, peer evaluation can serve as a major impetus for instructional and course improvement.

4. Administrators play a major role in faculty evaluation. At all institutional levels, they participate in personnel decisions as well as in decisions that affect the provision of faculty development opportunities.

In a survey of 1100 colleges and universities Gustad (1967) found that deans and chairmen were the two most frequently used sources of information in teaching evaluation. Clearly, the major responsibility for personnel and faculty development decisions lie with administrators. It is not so clear, however, on what data bases such decisions are made.

The absence of systematic evidence is a crucial factor in the deliberation of review committees. Hoyt (1977) found that many departments confused evaluation with ""eulogizing"; rather than making informed judgments, chairmen offered "pat them on the back" comments in evaluation sessions. The Southern Regional Education Board (1977) survey reported that casual

comments by students and faculty members were relied on very heavily in personnel decisions. Indeed, department chairmen often tended to give the benefit of the doubt to an instructor and give him or her an adequate rating unless overwhelming evidence to the contrary had come to their attention. Small wonder that a "general distrust of evaluation systems" prevailed among these institutions.

According to Genova et al. (1976), the administrator's role is a dual one --that of "direct observer" and that of "sifter" and "integrator" of evidence gathered from other sources of information. As direct observer, the administrator has an institutional responsibility in evaluation. Chairmen can observe a faculty member's performance in a variety of activities, and assess this performance in the light of departmental goals and specific role expectations. When administrators evaluate instruction, they could follow the same procedures proposed for colleagues. As sifter" and integrator" of information, the administrator could rely heavily upon student and colleague evaluation. In that role, the administrator's task is judiciously to combine information into a comprehensive view.

Perhaps the most important responsibility of an administrator is to facilitate the development and acceptance of a comprehensive system of instructional evaluation. Glasman (1976) referred to the administrator as "an identifiable leader" in instructional evaluation, because little would be pursued without such leadership. Institutions with systematic approaches to teaching evaluation surveyed in the Southern Regional Education Board (1977) study reported that better personnel decisions were being made following the implementation of such procedures and that faculty members were stimulated toward improved performance. Administrators can be very helpful to

instructors in identifying their strengths and weaknesses, and in supporting them in undertaking developmental activities.

5. A very important and direct method of evaluation is in the use of self-evaluation. Self-evaluation testifies to the autonomy and professional responsibility of the instructor. Without some form of self-appraisal, the quest for instructional improvement is doomed to failure.

The faculty member is the most appropriate person to describe the personal and professional context of his or her instructional role. An important aspect of self-evaluation is the annual report of professional activities -- courses taught, advisees, special projects, committee activities, and other relevant information. Such quantitative data provide an important context for the examination of qualitative information.

How valid are self-evaluations? As we recall from Snow White and the Seven Dwarfs, what we see in the "mirror on the wall" is not necessarily a true reflection of ourselves, but rather the image that we would like to see. Faculty members are no exception. That self-ratings correlate poorly with ratings of students, colleagues and administrators was shown by Blackburn and Clark (1975). These findings suggest that self-ratings would be of little value in tenure, promotion and salary decisions, but they could serve an important basis for instructional improvement. Growth and development can only occur where a recognition of the need for such improvement exists.

Research on Student Evaluation of Instruction

An extensive amount of research has examined various aspects of student evaluations of instruction. In this section, selected research findings will be

cited to address such issues as validity, reliability, influential factors, instrumentation, administration and impact of student evaluations.

1. How valid are student evaluations? Do the ratings measure what they purport to measure? In other words, do high ratings imply teaching effectiveness? While it may be true, as Marsh (1984) suggests, that no clearly defined criteria of teaching effectiveness can be agreed upon, a substantial amount of research helps to confirm the validity of student ratings. Among the measures used to establish the validity of student evaluations are achievement scores, ratings, self ratings, peer evaluations and ratings by trained observers or by means of videotape.

The studies by Sullivan and Shanes (1974) and Centra (1977) are of particular interest because in both studies students were randomly assigned to instructors. Through randomization, the differences in learning outcomes could not be attributable to differences between the students rather than differences in teaching (Leventhal et al., 1976). Both studies found substantial relationships between mean student ratings and student achievement.

In a meta-analysis that involved 41 independent studies, Cohen (1981) found strong support for student ratings as measures of teaching effectiveness. The average correlation on overall instructor rating and student achievement was .43, and between overall course rating and achievement was .47. Centra's (1977) research also showed that global ratings had higher validities than ratings of specific aspects of teaching.

Another approach to validate student ratings involved alumni. Some critics contend that students lack judgment and maturity during their undergraduate years; only some time after the university experience could former students make valid judgments of teaching effectiveness.

A comparison of student and alumni ratings has been used to establish the validity of student ratings. In a study that involved about 500 alumni, Centra (1974) found a high level of agreement between current student ratings of overall teaching effectiveness and similar ratings by alumni. The agreement was substantial, particularly at the extremes: very poor and very good teachers were identified as such by both groups.

A more recent study by Overall and Marsh (1980) compared responses from alumni who had rated a senior-level course with their ratings of the course and instructor a year following graduation. While the alumni ratings tended to be somewhat lower than ratings at the end of the course, the evaluations were highly correlated. Since it is clearly more difficult to obtain alumni surveys than current student ratings, and since the general agreement between them confirms the validity of student ratings, perhaps alumni surveys could be used primarily as information on course modifications.

Research studies have also consistently reported substantial agreement between student and colleague ratings (Maslow and Zimmerman, 1956; Hildebrand et al., 1971; Blackburn and Scott, 1975). As Batista (1976: 259) concluded, "There are activities and faculty behaviors that can be evaluated more validly and reliably by fellow colleagues than by anyone else." But classroom activities do not appear to be among them. Students have more exposure to their instructors teaching at the time they provide the ratings than do colleagues.

Although Dowell and Neal (1982) conclude from their extensive review of empirical studies that the validity of student ratings is modest at best and variable, they suggest that local validation of ratings should be performed. Different estimates of validity may be established for different departments, class levels, and course types.

A study of several measures of teaching effectiveness by Howard et al. (1985) found these methods more valid than prior findings would suggest. In particular, student ratings evidenced substantially greater validity coefficients than did self-report, colleague, and trained observer ratings. This finding is in general agreement with earlier research (Costin et al., 1971; McKeachie, 1979; Marsh, 1982), and provides strong support for the continued use of student ratings.

2. How reliable are student ratings? Reliability provides a measure of consistency or stability. Consistency may be established by comparing responses on odd-numbered items,with those on even-numbered items, or by comparing responses on the first half of the instrument with those on the second half. High, positive correlations between the two groups indicates a high degree of reliability. Stability may be derived from a comparison of results over time or over classes taught by the same instructor. Various comparisons can be made to yield a correlation coefficient as a measure of reliability.

A considerable number of studies reported by Doyle (1975) indicate that student ratings are very reliable in the sense of internal consistency. Using a sample of over 1000 students, Hildebrandt et al. (1971) reported reliability coefficients ranging in the high .80s. It should be noted, however, that internal consistency will be directly affected by the nature of the rating scale. If the scale measures more than one attribute or dimension of teaching effectiveness, reliability coefficients will drop accordingly. When subscales of a few items are constructed to form a composite scale, the internal consistency of these subscales may also drop markedly, largely because of the fewer number of items (Aleamoni and Spencer, 1973). The consistency of rating scores is also influenced by the type of item used in the rating scale. Thus, Crooks and Kane

(1981) found higher reliability on global measures than on specific or diagnostic measures of teaching effectiveness.

The stability of student ratings concerns itself with reliability over time or across different respondent groups. For six classes, McKeachie et al., (1978) found a correlation of .94 between mean student ratings at the end of a course and fifteen months later. High consistency in the students' global ratings of an instructor at different times in the same course were also reported by Centra (1979) and Crooks and Kane (1981).The correlation between mean ratings given by groups of students in the same class over time are very high, ranging from the mid .80s for classes of ten to the .90s for classes of twenty (Doyle, 1975).

A study by Gilmore, Kane and Naccarata (1978) showed that the use of ratings from five or more classes in which fifteen or more students responded provides a dependable measure of teaching effectiveness. They caution that the results should not be used for summative purposes, however, if the ratings were available from only one or two courses. For diagnostic purposes, the issue of reliability is clearly less important. As McKeachie (1979:393) concludes, "Reliability is not likely to be a concern for most uses of student ratings." Generally, student ratings are both highly consistent and stable measures of teaching effectiveness.

3. What factors influence student ratings? Virtually all discussions of instructional evaluation raise questions about factors that may influence student ratings. In general, Centra (1977) concludes that most extraneous factors have a relatively weak effect on student ratings. In the following paragraphs, research that examines the influence of student, teacher, and course characteristics will be highlighted.

A variety of student characteristics, such as age, gender, academic ability, grade point average, and college year, have weak or insignificant relationships with student ratings (Costin, Greenough and Mengers, 1971; Centra and Creech, 1976). Similarly, Abrami et al. (1982) found no meaningful or consistent relationship between students' personality characteristics and teacher ratings.

Perhaps the relationship that faculty members most often conjecture about is between student ratings and grades. Do students systematically give higher ratings to faculty members from whom they receive high grades? Or conversely, do faculty members award higher grades to students from whom they receive high ratings? If giving high grades means receiving high student ratings, then the "easy" instructor will be rated more favorably than the one who maintains "high standards" in grading. (Of course, it may be argued that students who do well may attribute their performance to the high quality of instruction received, and consequently evaluate their instructor accordingly.)

In his review of research studies, Doyle (1975) found mixed results. Some studies found no relationship between grades and ability; others found both positive and negative relationships. But even when relationships were found, they were only minimal.

Some consistent findings have been reported, however, on the relationship between grade expectations and student ratings. Feldman (1983) found that students in classes with higher expected grades gave their instructors higher ratings than did students in classes with lower expected grades. As Centra (1979) argues, the relationship between expected grades and ratings may provide modest evidence that students rate higher those courses in which they learn more. Similarly, McKeachie (1979) suggests that students who expect a

teacher to be good may be more attentive and highly motivated, and more likely to learn than those with poor expectations. Indeed, Leventhal et al. (1975) have shown that students may choose certain classes because of the instructor's reputation. Prior interest in a course generally is associated with higher ratings (Marsh and Cooper, 1981), but such a relationship probably reflects a high level of student interest and commitment to learning.

The relationship between teacher characteristics and student ratings has also been examined, but with few consistent results. In general, student ratings were not influenced by the instructor's age, gender, personality, teaching load or productivity (Bennett, Centra and Creech, 1976).

Neither were ratings influenced by tenure and rank. In a study of what happens to student ratings after instructors are promoted from one rank to another and, particularly, what happens when faculty members acquire both tenure and the highest academic rank, Eagle (1982) did not find support for the notion that student ratings declined once instructors had attained both tenure and the highest possible rank. Several studies (Brandenburg et al., 1978; Centra and Creech, 1976; Marsh, 1980) have shown, however, that professors receive higher ratings than do teaching assistants.

Years of teaching experience was found to relate to student ratings (Centra, 1979). Ratings of instructors increased during the first twelve years, and then declined somewhat. The slight decline in later years may be related to career shifts (some faculty members acquire substantial research or administrative responsibilities), but it may also reflect an attitudinal shift toward teaching. Highet (1976) suggests that instructors in their later years may assume too much knowledge on the part of students, particularly as their own

grasp of the subject matter has increased. Of course, some faculty members may simply become bored and indifferent in mid-career.

Finally, the relationship of student ratings and course characteristics have been examined. Since faculty members rarely control the conditions under which they teach, it is important to know how course characteristics influence student ratings.

Class size influences student ratings. Centra and Creech (1976) found that very small classes-- those with fewer than a dozen or so students-- usually are rated most highly, followed by students in classes with fifteen to thirty-five students. Students in very large classes (over 100) tend to give higher ratings than do those in classes with 35 to 100 students. Are instructors assigned to courses on the basis of teaching strengths? Do teachers prepare differently on the basis of class size? Perhaps the answers to these questions also suggest why discussion courses are generally rated higher than are lecture courses.

Students tend to give higher ratings to elective or general interest courses than to required courses (Feldman, 1976). Centra and Creech (1976) also found higher ratings for major than for college or university requirements. Similarly, students in upper level courses tend to give higher ratings than do students in lower level courses. Whitten and Umble (1980) conclude that an instructor can expect to receive higher student ratings if he or she teaches smaller instead of larger classes, non-core rather than core classes, and graduate in contrast to undergraduate classes.

There is also some evidence that the subject matter affects student ratings. Slightly higher ratings were found in the humanities than in the social sciences and natural sciences (Centra and Creech, 1976). These

differences may reflect the relative importance attached to teaching and research in these disciplines. As Landers (1976) suggests, most of these issues seem best to be handled through the development of appropriate norms and the assurance of enlightened interpretation of ratings. Unfortunately, factors affecting student ratings don't occur one at a time, allowing the precise isolation of the relative effect of each factor. In many settings, the interpreter simply assumes a random effect of the factors that influence student ratings.

4. What is the impact of student ratings? What is the research evidence that student ratings improve teaching? In a longitudinal analysis of the impact of student ratings, Gray and Brandenburg (1985) found only minor improvement over time. But earlier research provides more specific information about the impact of student evaluation of instruction.

Centra (1973b) designed an experimental study in which over 400 teachers at five different types of colleges were randomly assigned to one of three groups: a feedback or treatment group; a no-feedback or control group; and a post-test only group for comparative purposes. The major findings revealed that faculty members whose self-ratings were considerably higher than their student ratings, improved after only a half semester. A wider variety of instructors changed their instruction if given more than half a semester of time and if provided with information to help them interpret their scores. Cohen (1980) showed that instructors receiving mid-term feedback had significantly higher end-of-term ratings than those who did not receive feedback.

Hoyt and Howard (1978) found that faculty members who had been evaluated on the same course twice over a period of several years

showed significant gain scores. Pambookian (1974) found that teachers in the middle range of ratings tended to benefit more from feedback than did either those at the top or bottom of the ratings. Perhaps those with lowest ratings become discouraged. Centra (1973b) concluded that improvements were made most readily by instructors whose self-evaluations were considerably better than were their student ratings. Dissonance theory (Festinger, 1957) helps to explain the effort faculty members make to "close the gap" between student ratings and their own perceptions.

Students are generally lenient in their judgments about teaching. In her Nebraska study, Cross (1977) pointed out that students were typically generous in their ratings. Based upon a national sample of almost 200,000 students rating over 8000 classes between January 1976 and June 1978, Centra (1979) found that only about 12 percent of the teachers received less than average ratings; the remainder were average, good or excellent. These findings present a convincing argument for local or national norms. Comparing their ratings to those of other faculty members in the same department, faculty or university, can assist individual instructors in interpreting student ratings of their teaching.

It seems possible to increase the impact of student ratings by improved feedback. In a follow-up on a group of instructors who had gone through a consultation procedure in a previous semester, Erickson and Erickson (1978) reported that student ratings of these instructors improved. McKeachie (1979) reports a study at the University of Michigan that showed superior results for instructional evaluation after providing feedback of ratings by giving counselling to offer encouragement and to suggest alternative teaching strategies. Similar results were reported by Aleamoni

(1974) and Centra (1973a), indicating that instructors are more likely to improve if they have help in interpreting the feedback from student evaluations.

Little research has addressed the impact of student ratings on course selection. Do student ratings enable students to choose their courses more wisely? Although it is generally agreed that instructor reputation is a factor in student choice, the evidence is very sparse. Some student organizations collect and publish student ratings. Occasionally, neither the instrument nor the sampling appears to meet professional standards, thus leading faculty members to object to student evaluations. Kerlinger (1971) argues that student-sponsored ratings may result in instructor hostility, resentment and distrust. McKeachie (1979) reports a study that demonstrates that student ratings can influence student course selections. Borgida and Nisbett (1977), however, indicated that ratings have less effect on student choices than do face-to face consultations.

In a summary of student evaluations of university teaching, Marsh (1984) concluded that they seem to be useful for faculty members as feedback about their teaching, for students for use in course selection, and for administrators in making personnel decisions.

5. What kind of a rating scale should be used? Although this paper focuses upon the use of student ratings in instructional evaluation, clearly other procedures can be used to receive feedback from students. An instructor could invite student feedback periodically through informal dialogue with individual students or in classroom discussion. Some instructors occasionally invite a colleague or a peer consultant to attend class, and then to meet with students in the absence of the instructor to

ascertain how instructional improvement could be attained. Centra (1979) and Seldon (1984) provide helpful perspectives on ways to improve faculty effectiveness.

Many faculty members insist on developing their own rating scales. Teaching for them is a very private act, and academic freedom should protect instructors from the intrusion of external forces. The use of "standard" rating forms implies the existence of common elements in teaching effectiveness, and some faculty members assume that their teaching is so unique that it can only be evaluated with items constructed specifically for their courses. To be sure, some courses appear to be very different from most other courses, and all instructors have some unique teaching attributes. But considerable research evidence exists to substantiate common elements in effective teaching.

A number of studies have identified the qualities of instruction students believe to be related to superior teaching. From Feldman (1976) and Miller's (1974) reviews of these studies, the following qualities can be listed as being consistently reported: stimulation of interest, clarity, knowledge of subject matter, fairness, preparation, enthusiasm, friendliness, helpfulness, and openness to other's opinions.

The most common procedure used to identify major dimensions of teaching effectiveness is to factor analyze the results of student ratings. Factor analysis provides information on which items tend to cluster together and thus form a sub-scale or factor of the quality being measured. Feldman (1976) reviewed nearly 60 factor analytical studies of teaching effectiveness and identified three major clusters or dimensions: presentation of material, facilitation of learning, and regulation of students. Kulik and McKeachie

(1975) limited their review to eleven factor analytical studies and identified four dimensions as common factors: skill, rapport, structure and difficulty. Two other dimensions were described and interpreted by Isaacson et al. (1964): interaction and feedback. In his summary of factor analytical studies of student rating forms, Centra (1979) suggests the following dimensions or categories: organization, teacher-student interaction or rapport, teaching skill, workload, grading, impact on students and global or overall ratings.

In constructing a rating scale for conducting a comprehensive evaluation, it seems appropriate to select items from each of the major factors of teaching effectiveness. Berk (1979) recommends the use of at least two or three items for each dimension or factor to be measured. According to Centra (1979) and Braskamp and others (1984), specific diagnostic items that relate to the six major categories identified by Centra are recommended for use in instructional improvement due to their increased informational value. When instructors customize rating forms for instructional improvement, feedback provides information that is specific and relevant.

Global items, on the other hand, are the preferred type for personnel decisions. Cohen (1981) has shown that global items correlate more highly with student learning than do diagnostic ratings -- they are more valid measures of teaching effectiveness. The best way to arrive at an overall assessment of teaching effectiveness for summative purposes, according to Abrami (1985), is by averaging responses to several global items. McKeachie (1979) suggests that two to five general items probably provide sufficient feedback for making personnel decisions.

To construct a valid and reliable rating scale involves

considerable time and expertise. Usually, a large number of items are compiled from a variety of sources -- faculty, students, administrators, research literature and other forms. These items are categorized, using major dimensions of teaching effectiveness, and then analyzed for discriminatory value. Obvious duplication and "weak" items are eliminated through field testing and further analyses. Although various scaling approaches, have been used, Berk (1979) recommends a five-point, Likert-type scale from strongly disagree to strongly agree. Such a scale seems to have greatest utility in terms of construction and interpretability, and also seems to be highly reliable. About half of the items should be positively and the other half negatively worded to avoid acquiesence response. Rating scales have obvious strengths (Doyle, 1975); they are economical, readily summarized, and useful in making comparisons with earlier ratings or other instructor's ratings. Norms can serve as an incentive for instructional improvement and also as a basis for gross discriminations among faculty members.

A number of commercial student rating forms have been developed for instructional evaluation. Most popular among these are the Educational Testing Service's Student Instructional Report (Centra, 1979) and the Instructional Development and Effectiveness Assessment System (IDEA) from Kansas State University (Hoyt, 1973). The IDEA form contains three parts dealing, respectively, with instructional methods, student progress and course ratings. Both of these forms have national norms and the survey results can be interpreted comparatively within specified institutional groupings.

Two features that strengthen student rating forms and provide for

greater flexibility are instructor-specified items and open-ended items. Usually, written comments in response to open-ended items do not form a part of the "official" student evaluation data because of their subjective nature. Such comments are most helpful for diagnostic purposes when given directly to the instructor.

Another popular approach developed at Purdue University provides flexibility in item selection (Derry, et al., 1974). (Similar computerized forms have been developed at the University of Illinois, the University of Michigan, Ohio State and elsewhere.) The "cafeteria system" allows teachers to choose from an item catalog those items that seem most appropriate for evaluating a specific course. Usually, the items are classified into major descriptive and diagnostic categories, and a small number of core or global items that can be used for norming purposes.

Gray and Brandenburg (1985) found a difference between instructors who select items from an item catalog and those who use a standard form. Although the difference was not large, instructors who used the item catalog received higher ratings than those who did not. Since selecting items from an item catalog requires special processing by the instructor, they conjectured that possibly these instructors were more interested in teaching. They also found that instructors who used a departmental core of items scored higher than those who used a standard form. Clearly, a cafeteria system provides maximum flexibility and allows instructors to develop forms that address unique emphases in their teaching. At the same time, the inclusion of core or global items provides feedback that can readily be used for comparative purposes.

6. How should rating forms be administered? Variations in the

administration of rating forms can influence the results. To increase the validity and reliability of student ratings, special attention must be given to how, when, for what purpose and by whom student rating forms are administered.

There is some evidence that signed forms yield higher results than do anonymous forms (Stone et al., 1977). Feldman (1979) advises that rating forms should be kept anonymous in all cases, thus ensuring that a student cannot be penalized for giving low ratings.

The timing of student evaluations can have a direct effect upon ratings. Frey (1976) reports that ratings administered during final exams are generally lower than those given during the semester. The rating results may be unreliable if taken when students are under the pressure of examinations. How often evaluations are conducted may be a function of the purpose of the evaluation as well as the status of the instructors. If the ratings are designed for feedback for instructional improvement, Thorne (1980) recommends an evaluation should be given at least once every three times a course is offered. When a professor faces an administrative decision, at least five classes with a minimum of ten students should be evaluated (Centra, 1979). Tenured faculty members need not be evaluated annually; non-tenured faculty members should be evaluated in selected courses each semester.

Does the purpose of an evaluation influence the results? Feldman (1979) found some evidence that ratings are more positive if the stated use is for promotion than for improvement; Centra (1976) indicated that in a few instances students gave higher ratings when they understood the results would be used for administrative purposes.

There is clear evidence that results may be affected by who administers the evaluation. A study by Fentress and Swanson (1973) found that instructors received higher ratings by giving reinforcement and praise to the class at the time of the evaluation. Centra (1979) advises that a student or some type of a proctor distribute, collect, and place the rating forms into a sealed envelope, and that instructors not be present during the administration of the evaluation. As much as possible, student evaluations should be conducted in a prescribed and systematic manner, particularly when the results are used for summative purposes.

Recommendations for the Implementation of Student Evaluations

There is no one way to evaluate instruction. An overview of evaluation procedures clearly indicates that none by itself provides an adequate source of evaluative data. Universities would do well to develop comprehensive, multi-faceted systems of evaluation, encompassing a variety of evaluative procedures.

A discussion of student evaluation of instruction is incomplete without some recommendations for implementation strategies. Even the most ideal system of evaluation is of little value if faculty members do not accept it. Often faculty resistance stems from the manner in which a system is developed. Faculty members should participate fully in the design and implementation of instructional evaluation. And as much as possible, evaluation should be seen as an integral part of the reward system and growth opportunities within an institution (Menges, 1979).The following recommendations were derived from the research literature reviewed in this paper. Of particular significance in developing these recommendations for

the implementation of student evaluations, were selected writings on student evaluation (Braskamp et al., 1984; Centra, 1978; McKeachie, 1979; O'Hanlan, 1980; and Seldon, 1984).

1. A multi-faceted system of instructional evaluation should be adopted, including student, peer, administrator and self-evaluation activities. Following the adoption of procedures for student evaluation, other sources of evaluation should also be studied.

2. The evaluation system must be both credible and visible. Evaluation procedures should rest upon the wide-spread participation of faculty, students and administrators; campus newspapers and forums could be used to inform all campus constituencies about the evaluation procedures.When a campus-wide system is adopted,orientation and training sessions should be offered.

3. Student rating forms should be adopted or adapted to provide information for improved decision making in personnel matters as well as for instructional improvement. The primary purpose of instructional evaluation must be for increasing teaching effectiveness.

4. Standard procedures for conducting student ratings should be adopted. Usually, the evaluation should be conducted during the last two weeks of a course, but no later than two class periods before the end of the course. A student or faculty proctor should distribute, collect and deliver the ratings in a sealed envelope to a departmental or evaluation office for analysis. The evaluation results should not be reported to the instructor until the examination results have been submitted.

5. Avoid rating fatigue. The rating form could take about ten minutes of class time, but it should not be used for every class. Not every course should

be evaluated each time it is offered. New faculty members could be rated in at least one class each term, but tenured faculty members could ordinarily be evaluated only once every second or third term. New courses could be evaluated more frequently.

6. The student rating form should provide for flexibility by allowing instructors to develop optional items, and by inviting students to write comments in response to open-ended questions.

7. A small number of three to five global items could be used to measure the overall effectiveness of teaching. The responses on these items could be averaged and used as information in personnel decisions. Results from at least five classes of 15 students each should be available in order for data from student ratings to be included in summative evaluation. Student ratings should not be used in personnel decisions unless other data sources are also available.

8. Diagnostic items should be developed for each of the major dimension of teaching effectiveness and used to provide information to the instructor for teaching improvement. If a cafeteria system is adopted it may be appropriate to suggest the use of two or three items from each dimension in addition to the global items, unless a special interest is served by the evaluation. Indeed, it may be quite expedient to provide a "standard" set of items that could serve as a model form for student evaluation.

9. The confidentiality of data must at all times remain inviolate. Rating results are the property of the instructor, and they should only be shared with the consent and at the discretion of the assessed faculty member. Written comments by students should be made anonymously and provided directly to the instructor after the course is over.

10. Evaluation results must be carefully interpreted, particularly if they are used in personnel decisions. It may be helpful to offer training sessions in the interpretation of results. Whenever norms are compiled on items that involve sufficient responses, they should reflect the influence of extraneous factors on the evaluation results--department, class size, level and rank.

11. The administration must actively support the policies, procedures, and the basic costs of an evaluation system. All basic costs in the use of student ratings should be borne by central administration; additional research costs could be carried by the unit requesting such analyses.

12. There should be a clear linkage between evaluation and the reward system of the university. Incentives should encourage participation in evaluation, and resources should be provided to support instructional improvement, including the use of peer consultants.

13. Data storage security must be guaranteed. Cumulative rating results should be maintained, establishing local norms and addressing a variety of research questions, without the personal identification of any instructor. Follow-up studies could be undertaken to assess the effectiveness of the evaluation system and how to improve it.

14. Adequate time should be provided for the development, acceptance and implementation of student evaluation policies and procedures.The policies and procedures should be developed in consultation with an evaluation expert.

15.The student evaluation system should be staffed by an expert in instructional evaluation directly accountable to the academic vice-president, and the system should be monitored by a faculty-student advisory committee. The evaluation system should be "owned" by the faculty, not the administration.

References

Abrami, P.C. "Dimensions of Effective College Instruction." The Review of Higher Education. 1985, 8, 211-228.

Abrami, P. C., R. P. Perry and L. Leventhal. "The Relationship Between Student Personality Characteristics, Teacher Ratings, and Student Achievement," Journal of Educational Psychology. 1982, 74, 111-125.

Aleamoni, L.M. "Student Ratings of Instruction" in J. Millman (Ed.)Handbook of Teacher Evaluation Beverly Hills, CA: Sage, 1981, 110-145.

_____. The Usefulness of Student Evaluations in Improving College Teaching Teaching: University of Illinois, 1974.

Aleamoni, L.M. and R.E. Spencer. "The Illinois Course Evaluation Questionnaire: A Description of its Development and a Report on Some of its Results." Educational and Psychological Measurement. 1973, 33, 669-84.

Banz, M. L. and J. L. Rodgers. "Dimensions Underlying Student Ratings of Instruction: A Multidimensional Scaling Analysis." American Educational Research Journal. 1985, 22, 267 - 272.

Batista, E. E. "The Place of Colleague Evaluation in the Appraisal of College Teaching: A Review of the Literature." Research in Higher Education 1976, 4, 257-271.

Bennett, S.K. "Student Perceptions of and Expectations for Male and Female Instructors: Evidence relating to the Question of Gender Bias in Teaching Evaluation." Journal of Educational Psychology. 1982, 74, 170-179.

Benton, S. E. Rating College Teaching: Criterion Validity Studies of Student Evaluation-of-Instruction instruments. Washington, D.C.:AAHE-ERIC/ Higher Education Reports, No. 1, 1982.

Berk, Ronald, A. "The Construction of Rating Instruments for Faculty Evaluation: A Review of Methodological Issues." The Journal of Higher Education. 1979, 50, 650-669.

Blackburn, R.T., and Clark, M.J. "An Assessment of Faculty Performance:Some Correlates Between Administrators, Colleagues, Student, and Self-Ratings." Sociology of Education. 1975,48, 242-256.

Borgida, E. and R.E. Nisbett. "The Differential Impact of Abstract vs. Concrete Information on Decisions." Journal of Applied Social Psychology. 1977, 7, 258-271.

Brandenburg, D.C., J.A. Slinde and E.E. Batista. "Student Ratings of Instruction: Validity and Normative Interpretations." Journal of Research in Higher Education. 1977, 7, 67-78.

Braskamp, L.A. , D.C. Brandenburg and J.C. Ory. Evaluating Teaching Effectiveness. Beverly Hills: Sage, 1984.

Brown, D.L. "Faculty Ratings and Student Grades: A University-wide Multiple Regression Analysis." Journal of Educational Psychology. 1976, 68, 573-578.

Cashin, W.E. and B.M. Perrin. "Do College Teachers Who Voluntarily Have Courses Evaluated Receive Higher Student Ratings?" Journal of Educational Psychology. 1983, 75, 595-602.

Centra, J.A. "Effectiveness of Student Feedback in Modifying College Instruction." Journal of Educational Psychology. 1973a, 10, 395-401.

_____."The Relationship Between Student and Alumni Ratings of Teachers." Educational and Psychological Measurement. 1974, 34, 321-326.

_____. "Self-Ratings of College Teachers: A Comparison with Student Ratings." Journal of Educational Measurement. 1973b, 10 (4), 287-295.

_____. "Colleagues as Raters of Classroom Instruction." Journal of Higher Education. 1975, 46, 327-337.

_____. "The Influence of Different Directions on Student Ratings of Instruction." Journal of Educational Measurement. 1976, 13, 277-282.

_____. "Student Ratings of Instruction and Their Relationship to Student Learning." American Educational Research Journal. 1977, 14, 17-24.

_____. Determining Faculty Effectiveness. San Francisco, California: Jossey-Bass Publishers, 1979.

_____. "Evaluating Faculty Performance and Vitality" in W.R.Kirschling (Ed.) Using Student Assessments to Improve Performance and Vitality. New Directions for Institutional Research. San Francisco: Jossey-Bass, 1978, 5, 31-50.

Centra, J.A., and F.R. Creech. The Relationship Between Student, Teachers, and Course Characteristics and Student Ratings of Teacher Effectiveness. Project Report 76-1. Princeton, N.J.: Educational Testing Service, 1976.

Cohen, P.A. "Student Ratings of Instruction and Student Achievements: A Meta-analysis of Multisection Validity Studies." Review of Educational Research. 1981, 51, 281-309.

_____."Effectiveness of Student-Rating Feedback for Improving Instruction: A Meta-Analysis of Findings." Research in Higher Education. 1980, 13, 321-341.

Costin, F., W. T. Greenough, and R.J. Meneges, "Student Ratings of College Teaching: Reliability, Validity and Usefulness." Review of Educational Research, 1971, 41, 511-535.

Crooks, T.J. and M.J. Kane. The Generalizability of Student Ratings of Instructors: Item Specificity and Section Effects." Research in Higher Education, 1981, 15,305-313.

Cross, K. Patricia. "Not Can, But Will College Teaching be Improved" in J.A. Centra (Ed.) Renewing and Evaluating Teaching: New Directions for Higher Education, 1977, 1, 1-15.

Derry, J.O. "Can Students' Ratings of Instruction Serve Dual Purposes?" Journal of Higher Education, 1979, 50, 79-88.

Derry, J.O., W.F. Siebert, A.R. Starry, J.W. Van Horn and G.L. Wright.The Cafeteria System: A New Approach to Course and Instructor Evaluation. West Lafayette, Indiana: Purdue University, 1974.

Diamond, N., G. Sharp, and J. C.Ory, Improving Your Lecturing. Teaching: University of Illinois, 1978.

Donald, J.G. "The State of Research on University Teaching Effectiveness" in J.G. Donald and A.M. Sullivan (Eds.) Using Research to Improve Teaching and Learning. New Directions for Higher Education. San Francisco: Jossey-Bass,1985, 7-20.

Donald, Janet, G. and Bruce M. Shore. "Student Learning and the Evaluation of Teaching" in C.K. Knapper, G.L. Geis, C.E. Pascal and B.M. Shore (Eds.) If Teaching is Important . . . The Evaluation of Instruction in Higher Education. Toronto: Clarke, Irwin & Company, 1977.

Dowell, D.A. and J.A. Neal. "A Selective Review of the Validity of Student Ratings of Teaching." Journal of Higher Education, 1982, 53, 51-62.

Doyle, K.O. Jr. Student Evaluation of Instruction. Lexington: Heath and Company, 1975.

Dressel, Paul L. Handbook on Academic Evaluation. San Francisco: Jossey-Bass, 1976.

DuCette, J. and J. Kennedy. "Do Grading Standards Affect Student Evaluations of Teaching? Some new Evidence on an Old Question."Journal of Educational Psychology, 1982, 74, 308-314.

Eagle, Norman. "Student Ratings of Instruction Following Promotion in Professorial Rank." The Review of Higher Education, 1982, 5, 143-157.

Erickson, G.R. and B.L. Erickson. "Improving College Teaching: An Evaluation of a Teaching Consultation Procedure." Journal of Higher Education, 1979, 50, 670-683.

Feldman, K.A. "Consistency and Variability among College Students in Rating Their Teachers and Courses: A Review and Analysis." Research in Higher Education, 1977, 6, 233-274.

_____. "Course Characteristics and College Students' Ratings of their Teachers: What We Know and What We Don't." Research in Higher Education, 1978, 9, 199-242.

_____. "Grades and College Students' Evaluation of Their Courses and Teachers." Research in Higher Education, 1976, 4, 69-111.

_____."The Seniority and Instructional Experience of College Teachers as related to the Evaluation They Receive From Their Students."Research in Higher Education, 1983, 18, 3-124.

_____. "The Superior College Teacher from the Student's View." Research in Higher Education, 1976, 5, 243-288.

Festinger, L. A Theory of Cognitive Dissonance, Evanston: Row, Peterson, 1957.

Genova, W.J. et al. Mutual Benefit Evaluation of Faculty and Administrators in Higher Education, Cambridge, Mass: Ballinger,1976.

Gilmore, G.M., M.T.Kane, and R.W.Naccarato, "The Generalizability of Student Ratings of Instruction: Estimation of Teacher and Course Components." Journal of Educational Measurement, 1978, 15, 1-13.

Glasman, N.S. "Evaluation of Instructors in Higher Education." The Journal of Higher Education, 1976, 47, 309-326.

Gray, D.M. and D.C. Brandenburg. "Following Student Ratings Over Time With a Catalog-Based System." Research in Higher Education, 1985, 22, 155-168.

Greenwood, G.E. and H.J. Ramagli Jr. "Alternatives to Student Ratings of College Teaching." Journal of Higher Education, 1980, 51, 673-684.

Gustad, J.W. "Evaluation of Teaching Performance: Issues and Possibilities" in C. Lee (Ed.) Improving College Teaching, Washington: American Council on Education, 1967.

Highet, G. "The Need to Make it New." The Chronicle of Higher Education, 1976, 12, 40.

Hildebrand, M., R. C.Wilson, and E. R.Dienst. Evaluating University Teaching. Berkeley: University of California, 1971.

Howard, G.S., C.G. Conway and S.E. Maxwell. "Construct Validity of Measures of College Teaching Effectiveness." Journal of Educational Psychology, 1985, 77, 187-196.

Hoyt, D.P. "The Kansas State University Program for Assessing and Improving Instructional Effectiveness" in A. Sockloff (Ed.)Proceedings of the First Invitational Conference on Faculty Effectiveness as Evaluated by Students. Philadelphia: Temple University, 1973.

_____. "Background on the Uses and Misuses of Peer Evaluation -- Perspectives from Another Campus." Unpublished manuscript,Iowa City, Iowa, 1977.

Hoyt, D.P. and G.S. Howard. "The Evaluation of Faculty Development Programs". Research in Higher Education, 1978, 8, 25-38.
Isaacson, R.L. et al. "Dimensions of Student Evaluations of Teaching."Journal of Educational Psychology, 1964, 55, 344-351.

Knapper, C.K. "Improving Teacher Effectiveness." CAUT Bulletin, 1972, 21, 9-11.

Knapper, C.K. et al. If Teaching is Important . . . The Evaluation of Instruction in Higher Education. Toronto: Clarke, Irwin and Company, 1974.

Kulik, J.A. and W.J. McKeachie. "The Evaluation of Teachers in Higher Education" in F.N. Kerlinger (Ed.). Review of Research in Education (Volume 3). Itasca, Illinois: Peacock, 1975, 210-240.

Kulik, J.A. and C.C. Kulik. "Student Ratings of Instruction." Teaching of Psychology, 1974, 1, 51-57.

Leventhal, L., P.C. Abrami and R.P. Perry. "Do Teacher Rating Forms Reveal as Much About Students as About Teachers?" Journal of Educational Psychology, 1976, 68, 441-445.

Lin, Y., W.J. McKeachie and D.G. Tucker. "The Use of Student Ratings in Promotion Decisions." Journal of Higher Education, 1984, 55, 583-589.

Marsh, H.W. "Multidimensional Ratings of Teacher Effectiveness by Students from Different Academic Settings and Their Relationship to Student/Course/Instructor Characteristics." Journal of Educational Psychology, 1983, 75, 150-166.

_____. Students' Evaluation of University Teaching Dimensionality, Reliability, Validity, Potential Biases, and Utility." Journal of Educational Psychology, 1984, 76, 707-754.

_____. "The Influence of Student, Course, and Instructor Characteristics in Evaluations of University Teaching." American Education Research Journal, 1980, 17, 219-237.

_____. Validity of Students' Evaluations of College Teaching: A Multimethod Analysis." Journal of Educational Psychology, 1982, 74, 264-279.

Marsh, H.W. and T. Cooper. "Prior Subject Interest, Students' Evaluation, and Instructional Effectiveness." Multivariate Behavioral Research, 1981, 16, 82-104.

Maslow, A.H. and W. Zimmerman. "College Teaching Ability, Scholarly Activity, and Personality." Journal of Educational Psychology, 1956, 47, 185-189.

McIntyre, C. "Evaluation of College Teachers." Criteria, 1977, 6, 1-6. Ann Arbor: Center for Research on Learning and Teaching, University of Michigan.

McKeachie, W.J. "Student Ratings of Faculty: A Reprise." Academe, 1979, 65, 384-397.

McKeachie, W.J. and Y. G. Lin. "Multiple Discriminant Analysis of Student Ratings of College Teachers." Journal of Educational Research, 1975, 68, 300-305.

McKeachie, W.J., Y.G. Lin, and C.N. Mendelson. "A Small Study Assessing Teacher Effectiveness: Does Learning Last?" Contemporary Educational Psychology, 1978, 3, 352-357.

Meeth, L. Richard. "The Stateless Art of Teaching Evaluation " in Report on Teaching, Change Magazine, 1976, 2. New Rochelle, N.Y.

Menges, R.J. "Evaluating Teaching Effectiveness: What is the Proper Role for Students?" Liberal Education, 1979, 65, 356-370.

Miller, Richard I. Developing Programs for Faculty Evaluation: A Sourcebook for Higher Education, San Francisco: Jossey-Bass, 1974.

Moomaw, W. Edward. "Practices and Problems in Evaluating Instruction," in J.A. Centra (Ed.) Renewing and Evaluating Teaching: New Directions for Higher Education, 1977, 17, 77-91.

Murray, Harry, G. "Evaluation of University Teaching: Selective," in Newsletter, Ontario Universities Program for Instructional Development, 1980 (Feb/March), 2-9.

_____. Evaluating University Teaching: A Review of Research. Toronto: Ontario Confederation of University Faculty Associations, 1980.

O'Hanlon, J. and L. Mortensen. "Making Teacher Evaluation Work." Journal of Higher Education, 1980, 51, 664-672.

Overall, J.U. and H.W. Marsh. "Students' Evaluations of Instruction: A Longitudinal Study of Their Stability." Journal of Educational Psychology, 1980, 72, 321-325.

Pambookian, H.S. "The Initial Level of Student Evaluation of Instruction as a Source of Influence on Instructor Change after Feedback." Journal of Educational Psychology, 1974, 66, 52-56.

Perry, R.P., R.R. Niemi and K. Jones. "Effect of Prior Teaching Evaluations and Lecture Presentation on Ratings of Teaching Performance." Journal of Educational Psychology, 1974, 66, 851, 856.

Seldon, Peter. Changing Practices in Faculty Evaluation. San Francisco: Jossey-Bass, 1984.

_____. How Colleges Evaluate Professors. New York: Blythe-Pennington, Ltd., 1975.

_____. Survey of Current Faculty Evaluation Procedures. Report presented at the 4th International Conference on Improving University Teaching sponsored by the University of Maryland, 1978.

Shore, B.M. et al. Guide to the Teaching Dossier: Its Preparation and Its Use. Montreal: CAUT and McGill University Centre for University Teaching and Learning Services, 1980.

Smith, Albert B. Faculty Development and Evaluation in Higher Education. Washington, D.C.: The American Association for Higher Education, 1976.

Southern Regional Education Board. Faculty Evaluation for Improved learning. Atlanta, Georgia, 1977.

Stake, Robert E. "The Countenance of Educational Evaluation." Teachers College Record, 1967, 523-540.

Stufflebeam, Daniel E. "Evaluation as Enlightenment for Decision Making," Improving Educational Assessment and an Inventory of Measures of Affective Behavior. Washington: ASCD, NEA, 1969.

Sullivan, A.M. and G. R. Skanes. "Validity of Student Evaluation of Teaching and the Characteristics of Successful Instructors." Journal of Educational Psychology, 1974, 66, 584-590.

Sweeney, J.M. and A. F.Grasha. "Improving Teaching Through Faculty Development Triads." Improving College and University Teaching Yearbook. Corvallis: Oregon State University Press, 1978.

Thorne, G.L. "Student Rating of Instruction: From Scores to Administrative Decisions." Journal of Higher Education, 1980, 51, 207-214.

Whitten, B.J. and M.M. Umble. "The Relationship of Class Size, Class Level,and Core vs. Non-core Classification for a Class to Student Ratings of Faculty: Implications for Validity." Educational and Psychological Measurement, 1980, 40, 419-423.

Worthington, A.G. and P.T. Wong. "Earned and Assigned Grades on Student Evaluations of an Instructor." Journal of Educational Psychology, 1979, 71, 764-775.